Stat 202 - 2015S ~~S~~ - W~~11~~5 - ~~Tuesday~~ Tuesday (Pg 1)

Review

Before the exam we studied Sampling Distributions. The scenario we studied was the following

- We drew a sample of n individuals from a large population
- We measured a characteristic of each of the n individuals of the sample.
- The chacteristic could be
  - height or weight (quantitative)
  - or male / female
  - prefer red over green
  - or prefer green over red

  } categorical

For quantitative variables we sought to estimate the mean of the characteristic across the whole population, μ

For categorical variables we sought to estimate the proportion p of individuals with the characteristic across the whole population

In both cases we used only the sample to make the inference.

To estimate μ we used the sample mean $\bar{X}$
To estimate p we used the sample proportion $\hat{p}$

The sample mean $\bar{X}$ and sample proportion $\hat{p}$ were random variables, Every time we draw another sample we get different values for $\bar{X}$ and $\hat{p}$.

What sample we draw is random
$\bar{X}$ and $\hat{p}$ are numbers ~~[illegible]~~ that depend upon the outcome of the random phenomenon, (sample)
That's why they are random variables,

Main results of chapter 5 was the distribution of $\bar{X}$ and $\hat{p}$.

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \sigma/\sqrt{n}$$

approx $\left\{ \begin{array}{l} \mu_{\hat{p}} = p \\ \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} \end{array} \right\}$ When the binomial distribution is a good approximation

Good approximation when population is at least 20 x as large as sample.

Normal approximations

For large n $\bar{X} \overset{approx}{\sim} N(\mu, \sigma/\sqrt{n})$

If $np \geq 10$ and $n(1-p) \geq 10$

then $\hat{p} \overset{approx}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

New Chapter 6 Inference
For the rest of the semester
we are doing inference

purpose: draw conclusions from data

Formal inference emphasizes substantiating our conclusions via probability calculations,

Example: Are trees in a forest clustered or are they arranged randomly,

StatCrunch – longleaf1

Are these trees clustered or random and how do we find out?
584 trees old growth forest

Generate 584 random locations and see what it looks like;
4 colums seed 1

We already can see that most random forests look less clustered

But how do we be sure?

Key idea: Test Statistic

(Book sites another book which isn't in AU library and is $60
I want to find it on inter library loan
But I haven't done this yet)

So I am going to be vague.

My name for test stastistic

Clusterization Index.

Every forest has a clusterization index.

The higher the index the more ~~clustered~~ clustered the forest.

What foresth has the highest clusterization index?

All trees in same spot.

What forest has the lowest clusterization index,

Trees tiled regularly like a tree farm:

Random forests have clusterization indeces somewhere in between

Because they are random each random forest has a different clusterization index.

In particular the distribution of the test statistic is important.

Where does our old growth forest fall on that distribution

If it is ~~fall~~ into the tails of the distribution we can conclude the forest is clustered more than random If not, we ~~c~~ can't.

P-value

Who has heard of p-value?

The p-value is the probability that the test statistic would be as extreme or more extreme by random chance than what was observe in the data

In otherwords the ~~the~~ p-value is the probability that the clusterization index would be greater in a random forest than what was observed in the old growth forest,

More extreme in this case is greater,

In some cases you may be asking is the clusterization index less than (more extreme would be less than) And in others more extreme would be greater than a threshold and less than a different threshold For instance if you are asking is the forest random and you want to flag too clustered and not clustered enough, Called a two sided alternative

According to book the pvalue was 0.04

I don't know whether they used a one sided alternative or two.

$0.04 < 0.05$ ← traditional cutoff for significance

You would say the test is significant at 0.05 level (the traditional level),

You would reject the null hypothesis

Null hypothesis being that the forest is random.

You would conclude that the forest is not random.

Example #2

Researches want to know if a new drug is more effective than placebo

20 patients receive new drug
20 receive placebo

Twelve taking drug improve 60%
8 taking placebo improve

Stat → proportion stats → two sample
→ with summary

Sample 1
\# successes 12
\# obs 20

Sample 2
\# successes 8
\# obs 20

Null hypothesis is
Ho it is that both proportions are equal $p_1 = p_2$

Alternative hypothesis
do we care if its bigger $>$ less than $<$ or both $\neq$ ← two sided alternative

pvalue .20 large p-value

This means that if true proportions were the same,

In other words the probability of improving with the drug is the same as the probability of improving with the ~~the~~ placebo then the test statistic will be as extreme or more extreme 20% of the time

(I haven't told you got what the test statistic is)

Because this is so high we cannot conclude that the drug has an effect.

That doesn't mean it doesn't have an effect. It could mean that it doesn't have an effect it could also mean the sample size is too small to see the effect because the effect is also not large enough.

Review of terminology

test statistic
p-value
null hypothesis / alternative hypothesis
two sided alternative
versus one sided alternative

test statistic – a number you compute from data to be used to make an inference

~~p-value is the probability ...~~

Null hypothesis – ~~[illegible]~~ a statement that there is no effect – that the data observed ~~[illegible]~~ are caused by random chance
alternative hypothesis – statement that there is an effect
p-value – the probability assuming the null hypothesis is true that the test statistic will take on a value as extreme or more extreme as what is seen in data
One / two sided alternative – whether you consider extreme values less than or greater than only (one sided or both (two sided)